Similarity of the Neighbourhoods in Toronto and New York for Emigration Purposes

## Project Background

The aim of the project is to help people who are emigrating from Toronto to New York or vice versa to find similar neighbourhoods to where they currently live in terms of the venues nearby through visual aids. This will allow them to explore the different neighbourhoods and give them an idea of which neighbourhoods would be suitable for them to move to based on how similar they are to where they currently reside. It will also allow a general sense of how similar the two cities are and whether emigrating would be the right option for them.

The project will utilise Unsupervised Machine Learning using K-Means clustering to cluster similar neighbourhoods in both Toronto and New York based on the venues in their proximity.

## Data

The data to be used in the project consists of:

1. Toronto neighbourhood data including the Boroughs and Neighbourhoods from the following Wikipedia page: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
2. New York neighbourhood data including the Neighbourhoods and latitudes and longitudes from the following JSON file: <https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json>
3. Latitude and Longitude data from Geocoder or a CSV file to obtain the location data for each neighbourhood in Toronto
4. Foursquare data consisting of the nearby venues obtained using the Foursquare API with the latitude and longitude data

This data will be examined and explored using techniques such as visualisation before being pre-processed to convert it from it's raw form to a form which can be analysed. Any categorical variables will be encoded using One Hot Encoding. The data will be combined using the relevant criteria in order to allow the neighbourhoods to be compared and clustered based on their similarities in terms of venues. This will then be visualised to allow the analysis and learnings to be easily understood and evaluated to allow comparisons of neighbourhoods in both Toronto and New York.

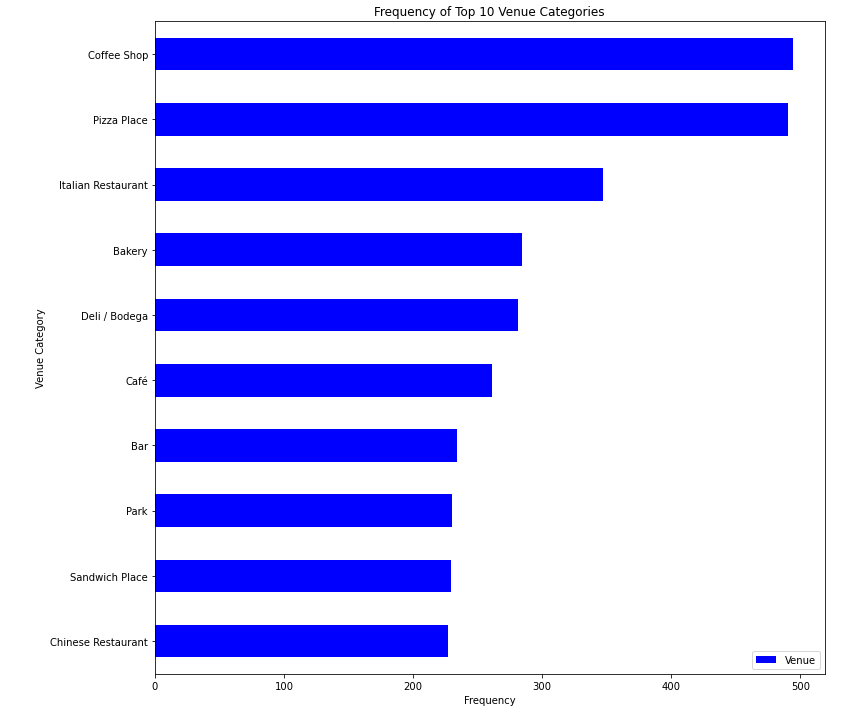
## Methodology and Implementation

The following methodology was implemented:

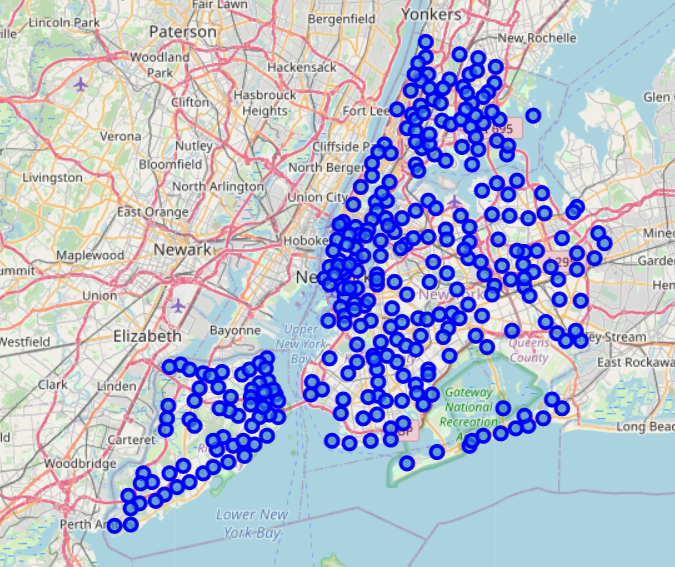
* Import the necessary libraries required for the analysis
* Gather the New York and Toronto data. The New York data is in the form of a JSON file on a server and the Toronto data is from a Wikipedia webpage
* Transform the JSON data which consists of nested dictionaries into a pandas data frame as this will make it easier to analyse
* Use Geopy to obtain the latitude and longitude of New York so it can be visualised
* Render the map of New York so it can be explored
* Scrape the Toronto data from the webpage into a pandas data frame
* Clean the data by removing any unassigned values
* Merge the data frame with the imported longitude and latitude values for the Toronto post codes
* Ensure the Toronto data frame matches the New York data frame
* Plot the Toronto map to visualise the data
* Combine the two data frames
* Invoke the Foursquare API to use the latitude and longitude data to obtain the nearby venues within a radius of 500 metres
* Group the neighbourhoods with the venues nearby
* Apply One Hot Encoding to allow the venue data from Foursquare to be used
* Obtain the mean frequency of the venue categories per neighbourhood
* Create a new data frame which lists the 10 most frequent venue categories per neighbourhood
* Run the K-Means Clustering algorithm on the data
* Visualise the results on a map
* Iterate the K-Means algorithm multiple times and observe the results to optimise the number of clusters
* Visualise the final data and define functions to allow users to find similar neighbourhoods based on the clusters

## Results

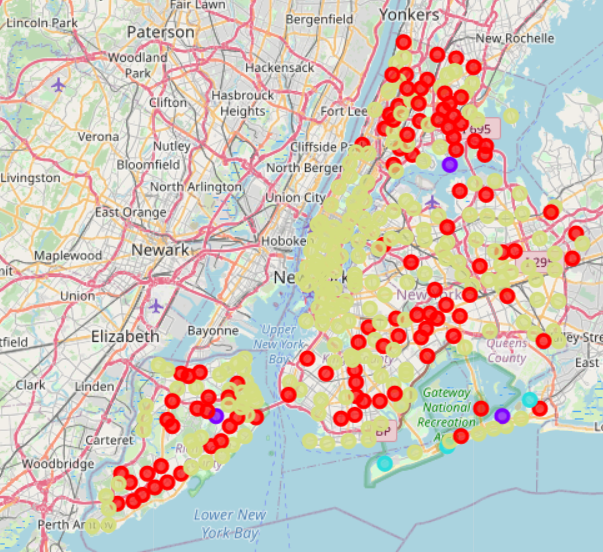
The Foursquare API data can be used to plot the top 10 most frequent venue categories across all of the neighbourhoods:



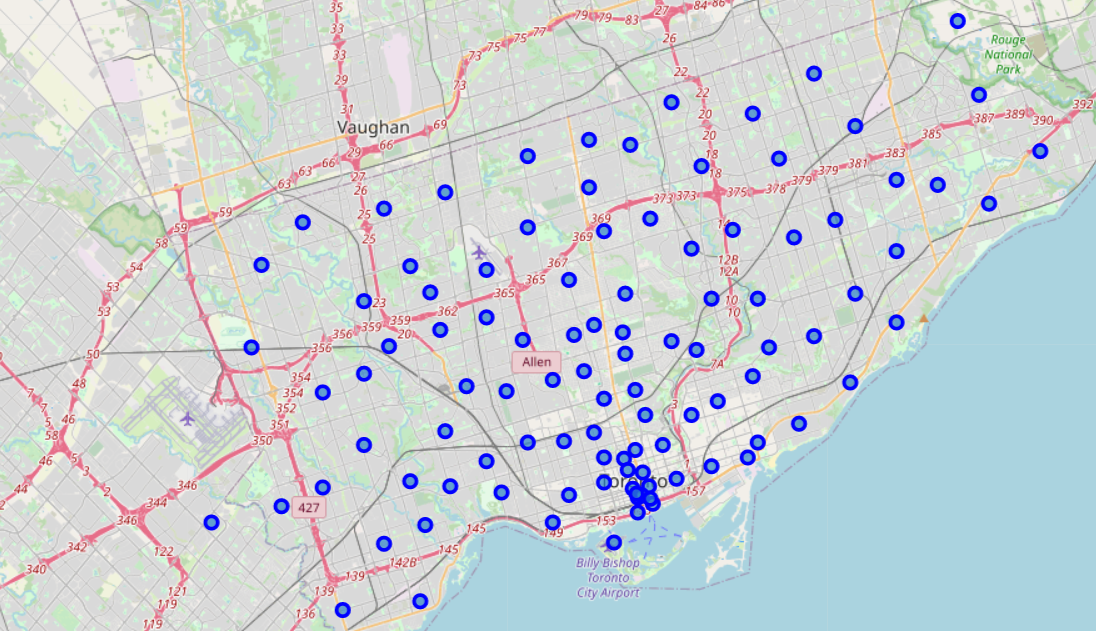
Map of New York before Clustering:



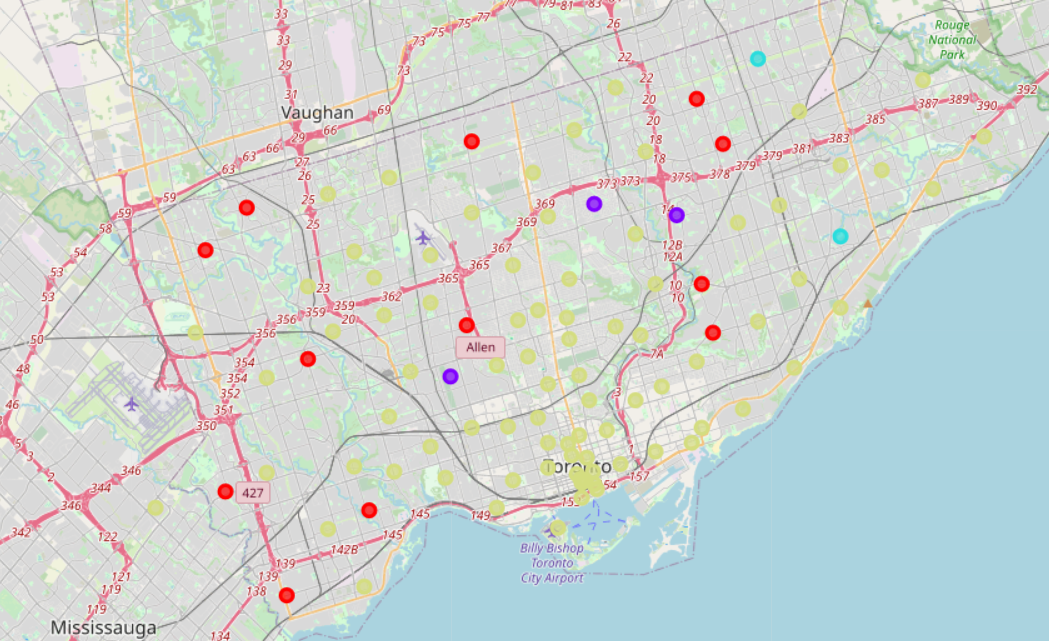
Map of New York after Clustering:



Map of Toronto before clustering:



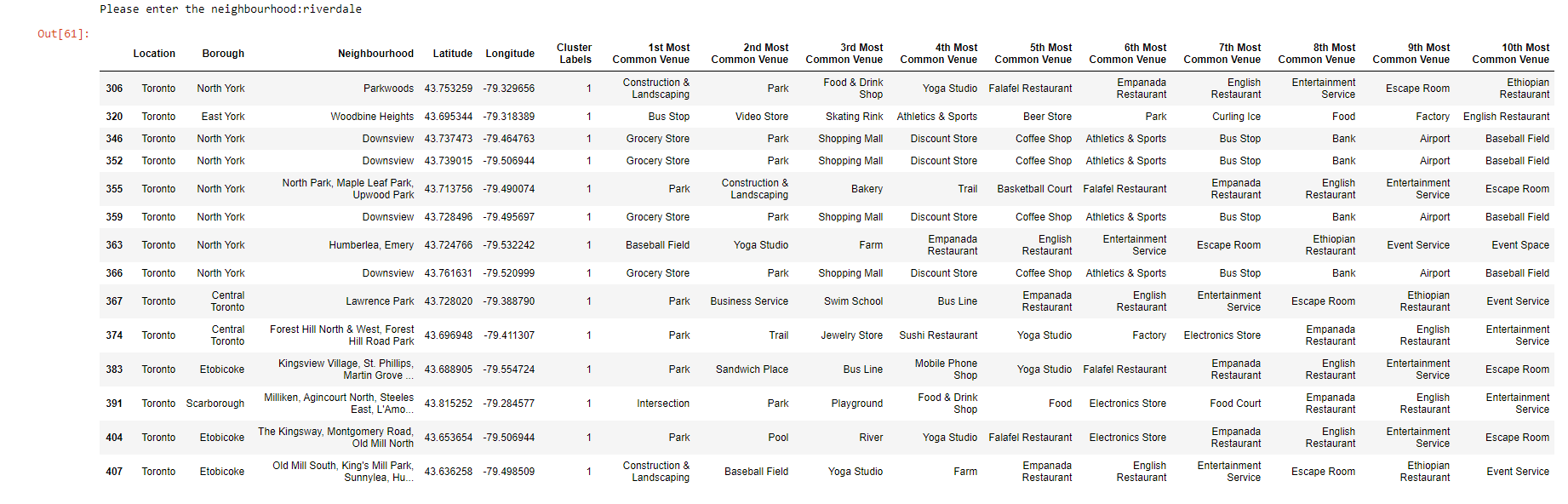
Map of Toronto after clustering:



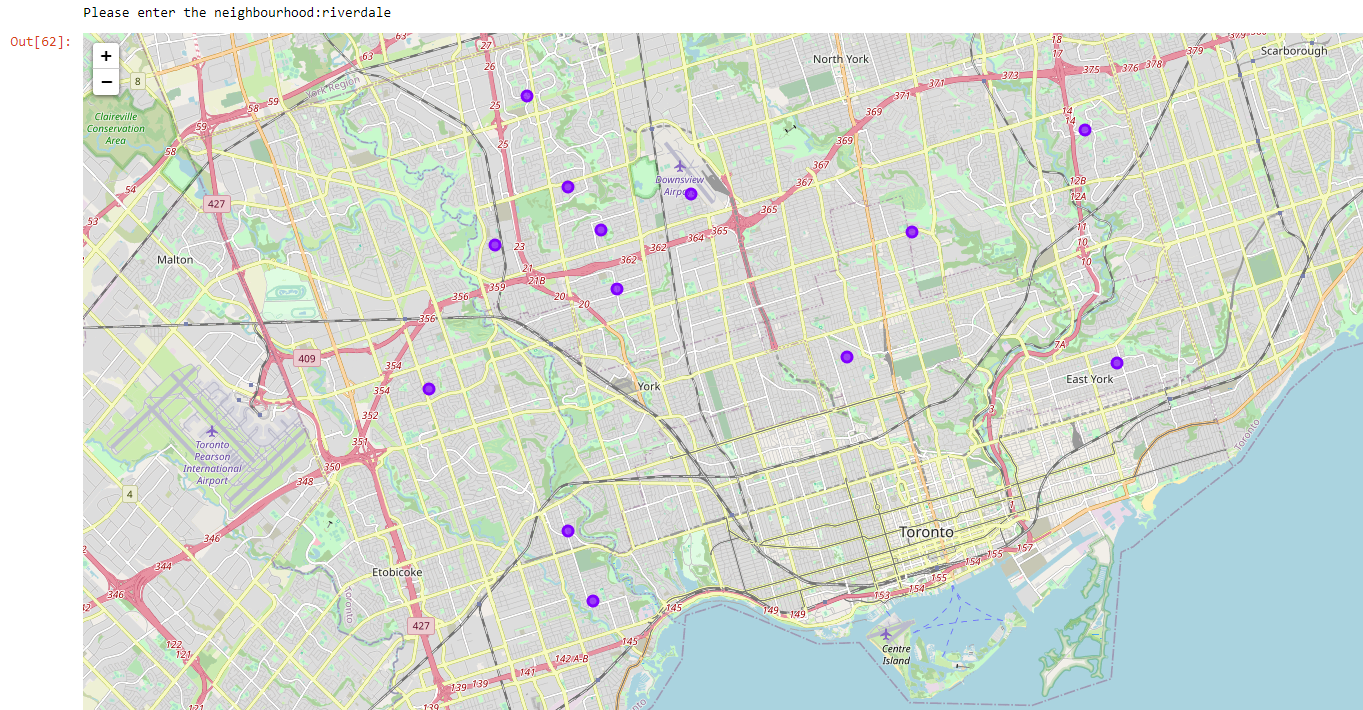
The algorithm was able to cluster similar neighbourhoods into the following broad categories:

* Cluster 1 - Restaurants and Convenience Stores
* Cluster 2 - Fitness and Nature
* Cluster 3 - Attractions and Landmarks
* Cluster 4 - Sports and Entertainment

Now that the neighbourhoods have been clustered, the data can be explored by the user. The user can enter the neighbourhood they currently reside in and a table will be shown of the similar neighbourhoods in the same cluster in the other location:



The user can also enter the neighbourhood they currently reside in and a map will be shown of the similar neighbourhoods in the same cluster in the other location:



## Discussion

I have implemented Unsupervised Machine Learning in the form of K-Means Clustering in order to group neighbourhoods across New York and Toronto according to their similarity in terms of venues nearby. I optimised the number of clusters by running the algorithm multiple times with differing numbers of clusters and observing the results before arriving at 4 clusters being a reasonable number. The data was visualised and explored before implementing the algorithm to ensure that it was suitable. This led to the pre-processing and refinement of the data to ensure the data could be utilised and any redundant or erroneous data was removed.

The clustering of the neighbourhoods into the 4 categories gives anyone thinking about emigrating between New York and Toronto an idea of which neighbourhoods may be suitable to explore further.

The algorithm can be further enhanced by providing additional data such as the average property prices and average income as well as the crime statistics of the neighbourhoods, which would allow for enhanced analysis and comparisons. This may lead to an increased number of clusters which would give the user an even better understanding of the neighbourhoods to explore further and assist in narrowing down the search for a suitable area.

## Conclusion

In conclusion, the data was successfully imported, pre-processed and K-Means Clustering implemented to cluster the neighbourhoods into 4 categories. Although the clustering was successful, there is scope for improvement and further analysis possibly using even more data such as financial data to help the user make an informed decision as to which neighbourhood to relocate to. Nevertheless, the analysis provides an adequate starting point for anybody interested in emigrating between New York and Toronto and allows the exploration and visualisation of the data in a quick and easy manner.